DOCUMENT RESUME

ED 223 673 TM 820 802

AUTHOR Woodcock, Richard W.
TITLE Interpretation of the Rasch Ability and Difficulty Scales for Educational Purposes.
PUB DATE 22 Mar 82
NOTE 12p.; Paper presented at the Annual Meeting of the National Council on Measurement in Education (New York, NY, March 20-22, 1982); Tables may not reproduce well due to small print.
PUB TYPE Speeches/Conference Papers (150) -- Reports - Research/Technical (143)

EDRS PRICE MF01/PC01 Plus Postage.
DESCRIPTORS Educational Diagnosis; Elementary Education; *Latent Trait Theory; *Scores; *Test Interpretation; *Test Use
IDENTIFIERS Keymath Diagnostic Arithmetic Test; *Rasch Model; Test Batteries; Woodcock Johnson Psycho Educational Battery; Woodcock Language Proficiency Battery; Woodcock Reading Mastery Test

ABSTRACT

Though many test developers have utilized item response theory in their work, few have taken advantage of the potential of item response theory for providing new interpretation procedures that accentuate the educational implications to be drawn from test scores. This paper describes several features, based upon the Rasch difficulty and ability scales, that have been incorporated into the interpretation schemes of test batteries currently used in the schools: the Woodcock Johnson Psycho-Educational Battery, Key Math Diagnostic Arithmetic Test, Woodcock Language Proficiency Battery, and the Woodcock Reading Mastery Tests. Among the features described are: common profiling of item difficulty and student ability, the "relative performance index," and the "instructional range." (Author)

****
* Reproductions supplied by EDRS are the best that can be made *
* from the original document. *
****

ED223673

INTERPRETATION OF THE RASCH ABILITY AND DIFFICULTY SCALES FOR EDUCATIONAL PURPOSES

Paper Presented to the Annual Meeting of the National Council on Measurement in Education, New York City, March 22, 1982

"PERMISSION TO REPRODUCE THIS MATERIAL HAS BEEN GRANTED BY

R. W. Woodcock

TO THE EDUCATIONAL RESOURCES INFORMATION CENTER (ERIC)."

Richard W. Woodcock
Measurement/Learning/Consultants
Tolovana Park, Oregon

U.S. DEPARTMENT OF EDUCATION
NATIONAL INSTITUTE OF EDUCATION
EDUCATIONAL RESOURCES INFORMATION CENTER (ERIC)

X This document has been reproduced as received from the person or organization originating it.
Minor changes have been made to improve reproduction quality.

• Points of view or opinions stated in this document do not necessarily represent official NIE position or policy.

Abstract

Though many test developers have utilized item response theory in their work, few have taken advantage of the potential of item response theory for providing new interpretation procedures that accentuate the educational implications to be drawn from test scores. This paper describes several features, based upon the Rasch difficulty and ability scales, that have been incorporated into the interpretation schemes of certain test batteries currently used in the schools. Among the features described are: common profiling of item difficulty and student ability, the "relative performance index" (RPI), and the "instructional range."

Introduction

The value derived from testing an individual is limited by the meaningfulness and interpretability of the scores which can be obtained. The development of the now traditional norm-referenced scores (including age scores, grade scores, percentile ranks, and standard scores) has made a significant contribution to the interpretation of test scores in education. On the other hand, the development of indexes of test performance that can describe the quality of an individual's test performance has been generally limited. A familar, but non-educational, example of such a criterion-referenced index is the Snellen index of visual acuity. The value 20/20 indicates that a person can see from a distance of 20 feet what the person with normal vision can see at 20 feet. An individual with 20/100 vision must be at 20 feet to see what the person with normal vision sees at 100 feet. The Snellen index, therefore, predicts the quality of an individual's visual performance under certain conditions.

Hearing loss, reported in dB's, is also an indicator of quality of performance. Normal hearing is referenced at zero on the dB scale. If an individual has a hearing loss of -40 dB, this indicates the additional amplitude necessary for that individual to hear a standard signal that a normal hearing person would hear at zero dB.

Students with hearing or visual problems are usually classified as handicapped, or in need of special services, on the basis of having significant deficits in their quality of visual or aural performance, not on the basis of falling below some point on a norm-referenced scale. On the other hand, in the area of mental retardation, classification is based primarily on a norm-referenced criterion such as "falling in the lower 3 percent of the general population." Thus, "mental retardation" is defined by statistical fiat rather than by the demonstration of some given degree of

performance deficit in cognitive functioning.

With the advent of item response theory and the emergence of practical analysis procedures such as the Rasch model, it is possible to devise ways of interpreting test scores that can add significantly to the utility of those scores. One outcome of such developments could be a shift toward using criterion-referenced indexes for classifying students into categories such as mentally retarded, gifted, or as having significant learning discrepancies. Such criteria would have better logical validity since they are based on a demonstrated deficit (or superiority) in quality of performance rather than on position in a norm group. A second outcome can be the provision of scales which predict the success or failure implications from placing a given student into instructional situations of various difficulty levels.

Though item response theory has opened new avenues for the interpretation of test scores, this potential seems to have been overlooked by many test developers. Information on some such interpretation schemes has been presented in Chapter 6 of the technical manual for the Woodcock-Johnson Psycho-Educational Battery (Woodcock, 1978). In addition, several suggestions, with a description of the necessary procedures, are presented in the chapter entitled "Choosing a Scale" in Best Test Design (Wright, 1979). The purposes of this paper are to bring such procedures to the attention of other test developers, to encourage broader adoption of those procedures, and to encourage continued exploration for more meaningful ways to interpret test results capitalizing on the principles of item response theory.

At least three interpretive schemes, based upon item response theory, have been incorporated into certain test batteries currently used in the schools. Those three procedures were developed in the context of preparing wide-age-range, individually-administered, supplied-response tests. There is little, if any, guessing factor with which to contend, and no problem of missing data. The item response model used in the analysis of data was the Rasch model. Any item which did not "fit the model" in respect to item discrimination characteristics was dropped from the item pool during test development. The calibration samples ranged from several hundred to several thousand subjects in size. The three applications of Rasch-based scores to test interpretation described in this paper include:

1. Common-scale profiling of difficulty and ability
2. The relative performance index (RPI)
3. The instructional range

Underlying Principles

A significant outcome from analyzing a set of item response data using the Rasch model is that person ability and item difficulty may be represented on a common scale. This concept is illustrated in Figure 1. The scale utilized in Figure 1 is a special $\log_9$ transformation of the Rasch scales (Woodcock & Dahl, 1971; Woodcock, 1978). The equation for transforming the Rasch ability and difficulty values into this scale has an additive constant of 500 and a multiplicative constant of 9.124. This transformation of the Rasch scales has particularly convenient probability values associated with certain points along the scale.

Figure 1 about here

If an individual is presented with a test item that has a difficulty on the same level as his or her ability, then there is a 50 percent chance that the individual will succeed. If the individual is presented with an item that is lower on the scale than his or her ability, then the probability of success is greater than 50 percent. If the item is above the individual's ability on the scale, the probability of success is less than 50 percent. Table 1 lists the probabilities of success associated with selected differences between a persons ability and item difficulty.

Table 1 about here

If desired, a probability value other than .50 may be used as a reference point for a special set of scores. For example, a pobability of .90 may be set as a reference level of success and then a subject's ability may be interpreted in respect to the success implications of presenting tasks which are more or less difficult than those which have a .90 probability of success. The Relative Performance Index (RPI), described later in this paper, is such an index. Table 2 lists the RPI values associated with a set of difference scores. This table has been reproduced from the Examiners Manual for the Woodcock-Johnson Psycho-Educational Battery (Woodcock, 1977).

Table 2 about here

Applications

Common-scale profiling of difficulty and ability. The most direct utilization of the Rasch scales is to take advantage of the fact that the difficulty values of items and the ability estimates for individuals can be expressed on a common scale. If a test consists of a set of items, each of which samples a significant subskill in a domain, then these items may be advantageously plotted along a common scale of difficulty and ability. This has been done for the diagnostic profile provided with the KeyMath Diagnostic Arithmetic Test (Connolly, Nachtman, & Pritchett, 1971). Figure 2 is a reproduction of the KeyMath diagnostic profile. There are 14 subtests listed vertically along the left side of the diagnostic profile. The items composing each subtest are positioned according to difficulty along the horizontal axis. Each item is based upon a behavioral objective stated in the manual and is identified by a mnemonic key for recognition of its content. The examiner uses a marking system to indicate which of the items were passed or failed. The row of numbers beneath each line represents the range of possible raw scores for that subtest. These raw scores have been positioned in respect to their corresponding Rasch ability value. Upon completion of testing, the examiner can observe an individual's pattern of subskill development, plotted by relative difficulty and correlated with the subject's overall ability. At the bottom of the profile is a scale of grade equivalents based upon the total test raw score.

Figure 2 about here

An extension of the concept of the diagnostic profile, though it was not

incorporated into the KeyMath diagnostic profile, would be the provision of a moveable scale to facilitate predicting the likelihood of success on items (or on the real-life tasks to which item performance can be generalized) falling at various distances above or below an individual's ability. Any item with a difficulty value corresponding to the person's ability has a 50 percent chance of success. On the actual KeyMath profile (not on the reduced reproduction in Figure 2) any item which falls exactly 5/6 of an inch above the person's ability has a 25 percent probability of success for that individual. Any item which falls twice that far above the person's ability has a 10 percent chance of being passed, and any item falling 3 times that distance has a 4 percent chance. On the other hand, items which fall 5/6 of an inch below the person's ability have a 75 percent chance of success. Those items falling twice that far have a 90 percent chance of success, and those falling 3 times that far have a 96 percent chance of success. The innovative test developer could certainly capitalize on the use of a moveable scale to enhance the interpretation of a new criterion-referenced or diagnostic skills test.

Relative performance index (RPI). The RPI is an index that has been developed to provide a predictive statement about a person's expected quality of performance relative to some standard performance. RPIs have been utilized thus far in four test batteries (Woodcock, 1973; Woodcock & Johnson, 1977; Woodcock, 1980; Woodcock, 1981). The RPI is similar to the Snellen index and the interpretation of the two indexes is quite similar (however, the constant term in the Snellen index is usually the numerator and the constant term in the RPI is the denominator).

An example of an RPI is "80/90". If this index were obtained from a reading test, the value would indicate that the subject is predicted to perform with 80 percent success those reading tasks that average students in a comparison group perform with 90 percent success. The most frequently used comparison is the subject's grade placement. An RPI of 32/90, based on the student's grade level would mean that the student is predicted to perform with 32 percent success those tasks that average students in that grade perform with 90 percent mastery. An RPI of 99/90 means the subject is predicted to perform with 99 percent success those tasks performed with 90 percent success by average students in that grade.

The RPI may be based on age instead of grade placement. For example, in a measure of oral language development used with preschoolers, it might be desired to predict the child's quality of oral language performance relative to others of the same age. A 5-year-old with an RPI of 79/90 would be one who has demonstrated the ability to perform with 79 percent mastery those language tasks that his or her average agemates perform with 90 percent mastery.

The comparison in an RPI does not need to be extra-individual but can be intra-individual. In the area of learning disabilities, for example, clinicians frequently compare an individual's actual achievement test score to an "expected" achievement test score. In the section of the Woodcock-Johnson where actual achievement is compared to expected achievement in order to determine the presence of an achievement discrepancy, RPIs indicate the student's percent of success on tasks that were predicted to be 90 percent successful based on the student's expected achievement score. In this context, an RPI of 40/90 in mathematics would indicate that the student is achieving with only 40 percent success those mathematics tasks for which the expectancy score predicted 90 percent success.

Another example of the use of an RPI based on intra-individual differences is found in the Woodcock Language Proficiency Battery (1980, 1981). RPIs are used here to report the level of success on language tasks in the subject's non-dominant

language when given language tasks which can be performed with 90 percent success in the dominant language. Thus, a bilingual student who has been tested in both Spanish and English and whose dominant language is Spanish might have an RPI of 15/90. This means that given the language tasks that can be performed by this individual with 90 percent success in Spanish, the same language tasks can be performed with only 15 percent success in English. This comparison is possible as the result of a special Rasch-based linking study conducted during the norming of the Spanish language battery (Woodcock, 1981).

RPIs are of particular value for two reasons. First, they are statements that describe relative quality of performance, not relative standing in a group. Second, if the population, on the average, becomes better or poorer over the years, relative performance scores still have the same meaning (just as does the Snellen index). On the other hand, norm-referenced scores, such as percentile ranks, are statements describing standing in a group. Standing at the fifth percentile may represent a quite different degree of performance deficit if the average level, or the distribution of that ability in the population shifts, from one norming to the next. Because of this constancy in meaning, indexes of relative performance such as the RPI, could have considerable utility in definitions of conditions such as mental retardation or giftedness, just as the Snellen index has been so valuable in the field of visual impairment.

Instructional range. The purpose of the instructional range is to provide a grade scale guide to the range of instruction that would be perceived as "easy" by a given subject (RPI=96/90) to the level that would be perceived as "difficult" or frustrating (RPI=75/90). The instructional range was introduced in the Woodcock Reading Mastery Tests (1973). The interpretation of this range directly parallels observing the difference between the "easy reading level" and the "frustration level" obtained by the informal reading inventory procedure (IRI) as described by Betts (1957). Thus, the instructional range based on an objective test, can present certain instructional implications in the manner of an informal reading inventory but with the added benefits of a standardized presentation and more precise (reliable) measurement. In subsequently developed test batteries, the instructional range concept has been extended from the interpretation of reading test scores to include oral language, written language, mathematics, and knowledge scores.

Figure 3 illustrates the application of the instructional range concept to a battery of individually administered achievement tests (Woodcock & Johnson, 1977). The "X" within each instructional range band marks the level at which it is predicted that instruction would be perceived by the student as "average" in difficulty (RPI = 90/90). The vertical line represents the student's grade placement and the RPIs (based on grade) indicate the level of success which is predicted for this student if assigned tasks that his average grademates would perform with 90 percent success. For example, the instructional range for reading achievement in Figure 3 indicates that reading instruction or assignments which are of average difficulty for grade 1.8 students would be perceived as easy by this boy (RPI = 96/90). Reading tasks at the grade 2.2 level would be perceived as average in difficulty (RPI = 90/90) while reading tasks at the 2.6 level would be perceived as difficult and frustrating (RPI = 75/90). If this student were presented reading tasks that his average grademates could perform with 90% success, we would predict a mere 4 percent level of success for him (a traumatic experience!). The instructional range, and the instructional implications profile, have been especially developed to communicate to teachers and others the performance implications of placing a student at various difficulty levels of instruction.

Figure 3 about here

The results of a validity study designed to evaluate the predictive capabilities of Rasch-based scores is reported in the manual for the Woodcock Reading Mastery Tests (Woodcock, 1973). The purpose of that study was to demonstrate whether the instructional range from that battery could provide valid information for predicting which of several specified items from another pool of items (administered one week later) would be performed at the "easy reading," "reading grade," and "failure reading" levels. Note that the requirement in this study was that a score, determined by the subject's 50 percent level of functioning, was used to predict which items from a pool of criterion tasks would be performed at the 75 percent, 90 percent, or 96 percent level of success. The data for this study were obtained from 103 students in the second grade and 102 students in the seventh grade. The results are presented in Table 3. A chi-square test of significance indicated that there was no significant difference between the predicted and observed levels of performance on the criterion tasks for each of the five skill areas and for Total Reading as measured by the WRMT. These results provide support for the validity of using a subject's Rasch-based score to predict specified levels of success with tasks of graded difficulty which do not necessarily coincide with the subject's ability level.

Table 3 about here

A clinical study by Memory, Powell, and Callaway (1980) compared scores from the Woodcock Reading Mastery Tests with IRI and Spache instructional levels. The subjects were 62 children in a reading clinic population at the University of Georgia. The investigators report high correlations and small difference scores between the Woodcock scores and the IRI instructional levels based upon the Betts criteria of 95 percent word recognition accuracy and 75 percent comprehension. They conclude, "based upon a comparison between the performance of 62 children on two sets of graded passages using the Betts instructional level criteria and their performance on the Woodcock, this battery of tests appears to be an accurate alternative for assessing the abilities measured by an Informal Reading Inventory."

Conclusion

The potential for elaborating the interpretation of test scores based on the Rasch model seems analogous to walking out the end of a tunnel. Back in the tunnel it was possible to see but one point on the horizon. Upon emerging from the tunnel, the horizon is there to be surveyed. Traditional test scores allow us to predict performance only at the point represented by the obtained test score. Properly developed Rasch-based scores allow the user to predict performance on any task given its scaled distance from the individual's ability score. This principle provides a powerful foundation for the development and use of interpretation procedures that can accentuate the implications to be drawn from educational test scores.

REFERENCES

Connolly, A. J., Nachtman, W., & Pritchett, E. M. KeyMath Diagnostic Arithmetic Test. Circle Pines, Minnesota: American Guidance Service, 1971, 1976.

Memory, D., Powell, G., & Callaway, B. "A Study of the Assessment Characteristics of Woodcock Reading Mastery Tests." Reading Improvement, Spring 1980, No. 1, 48-52.

Wright, B. D., & Stone, M. H. Best Test Design. Chicago: Mesa Press, 1979.

Woodcock, R. W. Woodcock Reading Mastery Tests. Circle Pines, Minnesota: American Guidance Service, 1973.

Woodcock, R. W. Development and Standardization of the Woodcock-Johnson Psycho-Educational Battery. Hingham, Massachusetts: Teaching Resources, 1978.

Woodcock, R. W. Woodcock Language Proficiency Battery--English Form. Hingham, Massachusetts: Teaching Resources, 1980.

Woodcock, R. W. Bateria Woodcock Psico-Educativa en Espanol. Hingham, Massachusetts: Teaching Resources, 1981.

Woodcock, R. W. & Johnson, M. B. Woodcock-Johnson Psycho-Educational Battery. Hingham, Massachusetts: Teaching Resources, 1977.

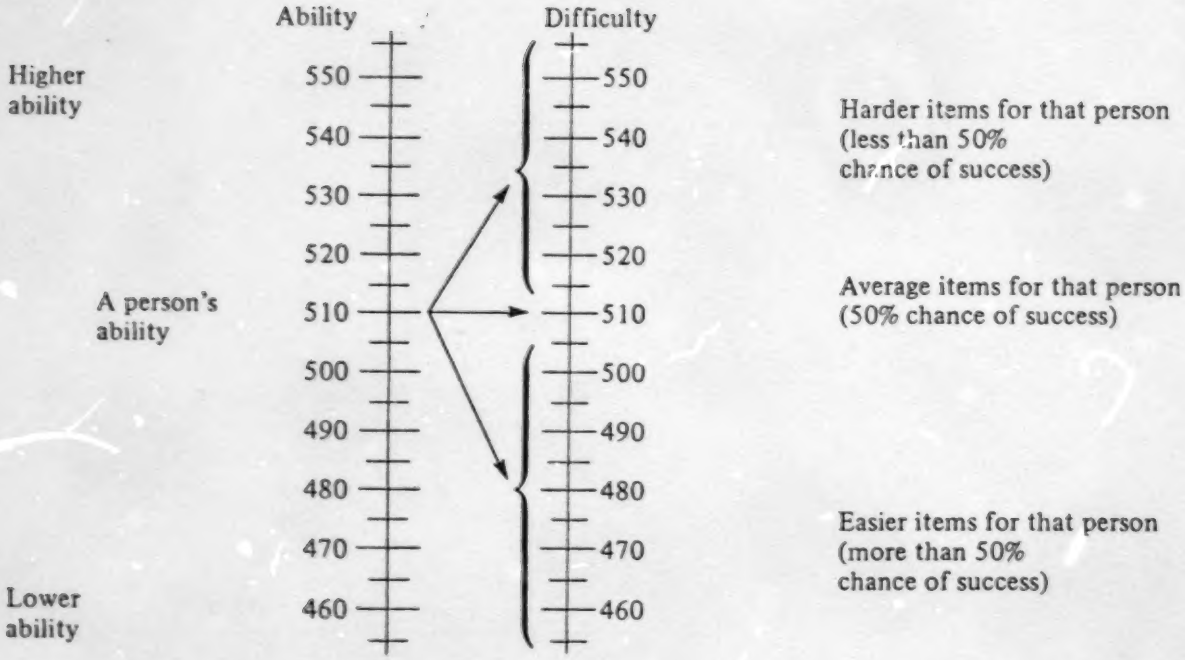

Figure 1. Relationship between a person's ability and task difficulty.

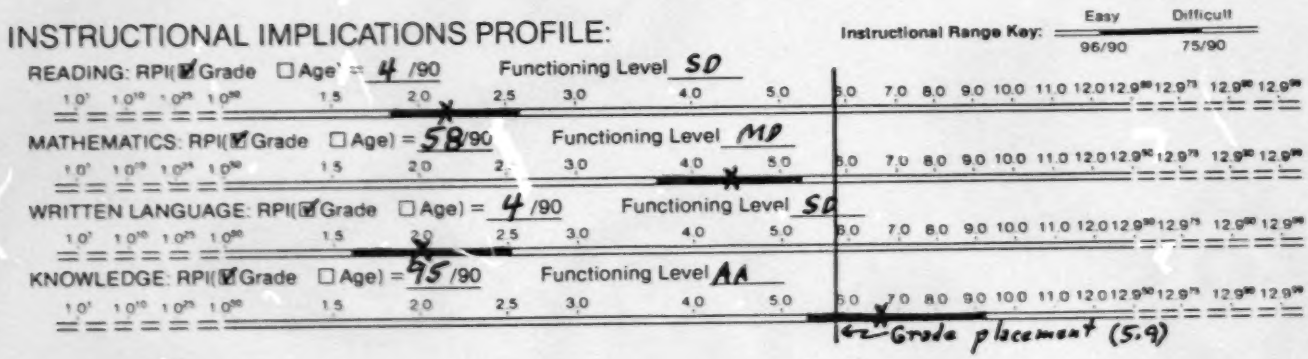
INSTRUCTIONAL IMPLICATIONS PROFILE:
Instructional Range Key: Easy 96/90 Difficult 75/90
READING: RPI(☑Grade ☐Age) = 4 /90 Functioning Level SD
MATHEMATICS: RPI(☑Grade ☐Age) = 58/90 Functioning Level MD
WRITTEN LANGUAGE: RPI(☑Grade ☐Age) = 4 /90 Functioning Level SD
KNOWLEDGE: RPI(☑Grade ☐Age) = 95 /90 Functioning Level AA
Grade placement (5.9)

Figure 3. Profile of instructional ranges for a fifth-grade boy referred for school learning problems (from the Woodcock-Johnson Psycho-Educational Battery).

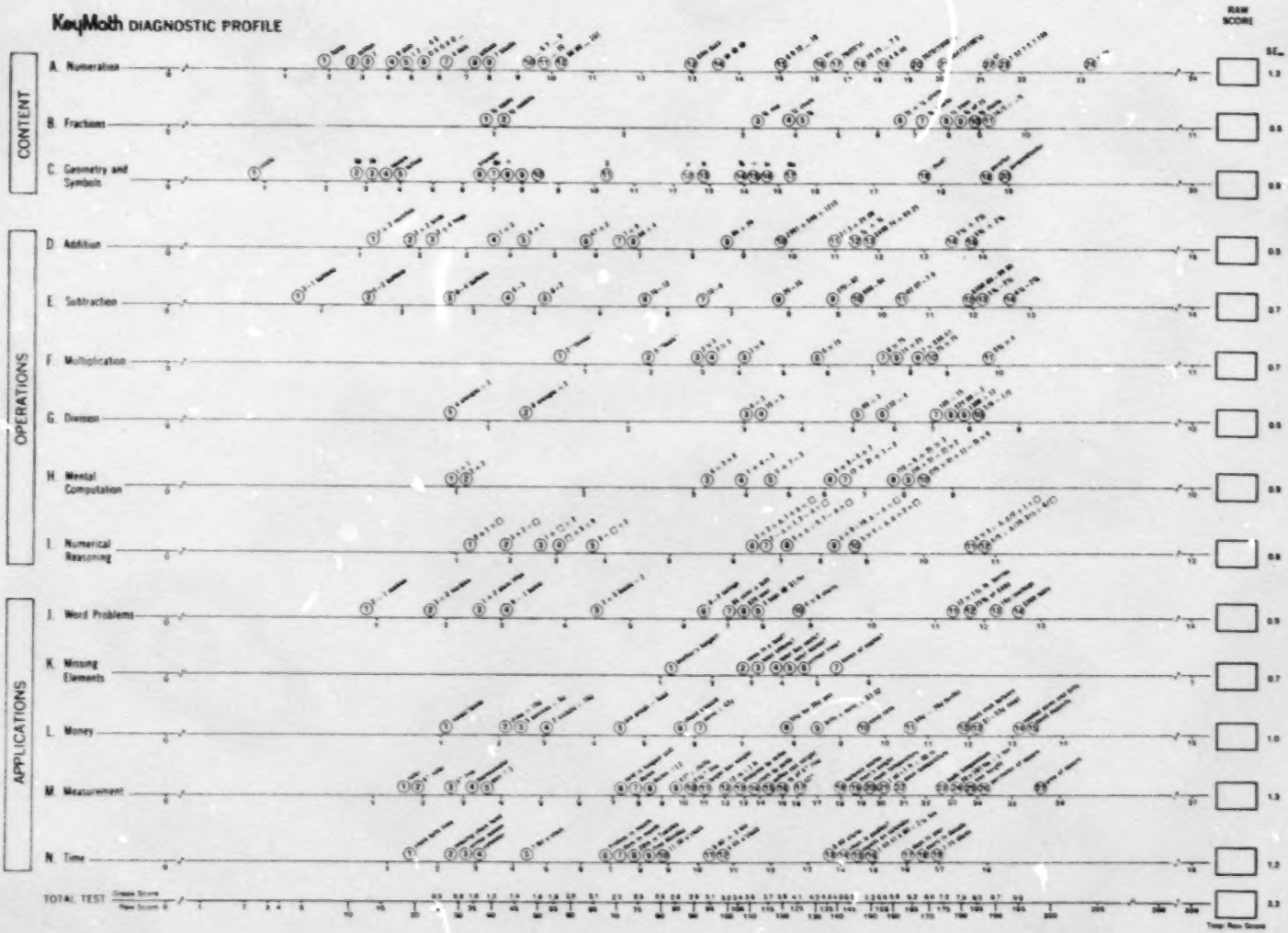

Figure 2. KeyMath diagnostic profile (reproduced by permission of American Guidance Service).

Table 1. Probability of Success Given the Difference on the Scale Between Ability and Difficulty

ABILITY MINUS DIFFICULTY (W_{A-D})	PROBABILITY OF SUCCESS (P)	ABILITY MINUS DIFFICULTY (W_{A-D})	PROBABILITY OF SUCCESS (P)
+50	.996	0	.500
+45	.993	- 5	.366
+40	.988	-10	.250
+35	.979	-15	.161
+30	.964	-20	.100
+25	.940	-25	.060
+20	.900	-30	.036
+15	.839	-35	.021
+10	.750	-40	.012
+ 5	.634	-45	.007
0	.500	-50	.004

TABLE 3. RESULTS OF THE INSTRUCTIONAL RANGE PREDICTIVE STUDY

TEST	PREDICTED PERCENT CORRECT			DEGREES OF FREEDOM	χ^2	PROBABILITY
	96.0	90.0	75.0			
Letter Identification Observed percent correct.	96.1 (n = 615)	92.7 (n = 423)	78.8 (n = 321)	2	0.92	>.50
Word Identification Observed percent correct.	95.6 (n = 612)	88.0 (n = 612)	71.8 (n = 610)	2	1.15	>.50
Word Attack Observed percent correct.	92.5 (n = 308)	86.0 (n = 399)	77.4 (n = 452)	2	1.48	>.30
Word Comprehension Observed percent correct.	93.5 (n = 495)	85.8 (n = 544)	74.7 (n = 593)	2	1.39	>.50
Passage Comprehension Observed percent correct.	96.6 (n = 494)	91.0 (n = 577)	73.6 (n = 602)	2	0.27	>.80
TOTAL READING Mean percent correct....	94.9	88.7	75.3	8	5.21	>.70

TABLE 2 – RELATIVE PERFORMANCE INDEX (RPI) AND FUNCTIONING LEVEL

CLUSTER DIFFERENCE SCORE	RPI	Functioning Level[1]
29 and above	100[2]/90	VS
28	99/90	VS
27	99/90	VS
26	99/90	VS
25	99/90	S
24	99/90	S
23	99/90	S
22	99/90	S
21	99/90	S
20	99/90	S
19	99/90	S
18	98/90	S
17	98/90	S
16	98/90	S
15	98/90	AA
14	98/90	AA
13	97/90	AA
12	97/90	AA
11	97/90	AA
10	96/90	AA
9	96/90	AA
8	96/90	AA
7	95/90	AA
6	95/90	AA
5	94/90	A
4	93/90	A
3	93/90	A
2	92/90	A
1	91/90	A
0	90/90	A
- 1	89/90	A
- 2	88/90	A
- 3	87/90	A
- 4	85/90	A
- 5	84/90	A
- 6	82/90	BA
- 7	81/90	BA
- 8	79/90	BA
- 9	77/90	BA
- 10	75/90	BA
- 11	73/90	BA
- 12	71/90	BA
- 13	68/90	BA
- 14	66/90	BA
- 15	63/90	BA
- 16	61/90	MD
- 17	58/90	MD
- 18	55/90	MD
- 19	53/90	MD
- 20	50/90	MD
- 21	47/90	MD
- 22	45/90	MD
- 23	42/90	MD
- 24	39/90	MD
- 25	37/90	MD
- 26	34/90	SD
- 27	32/90	SD
- 28	29/90	SD
- 29	27/90	SD
- 30	25/90	SD
- 31	23/90	SD
- 32	21/90	SD
- 33	19/90	SD
- 34	18/90	SD
- 35	16/90	SD
- 36	15/90	SD
- 37	13/90	SD
- 38	12/90	SD
- 39	11/90	SD
- 40	10/90	SD
- 41	9/90	SD
- 42	8/90	SD
- 43	7/90	SD
- 44	7/90	SD
- 45	6/90	SD
- 46	5/90	SD
- 47	5/90	SD
- 48	4/90	SD
- 49	4/90	SD
- 50	4/90	SD
- 51	3/90	SD
- 52	3/90	SD
- 53	3/90	SD
- 54	2/90	SD
- 55	2/90	SD
- 56	2/90	SD
- 57	2/90	SD
- 58	2/90	SD
- 59	1/90	SD
- 60	1/90	SD
- 61	1/90	SD
- 62	1/90	SD
- 63	1/90	SD
- 64	1/90	SD
- 65	1/90	SD
- 66	1/90	SD
- 67	1/90	SD
- 68	1/90	SD
- 69 and below	0[3]/90	SD

[1] VS–Very Superior
S–Superior
AA–Above Average
A–Average
BA–Below Average
MD–Moderate Deficit
SD–Severe Deficit

[2] Approximate value (greater than 99.5)

[3] Approximate value (less than 0.5)